Applications of Artificial Intelligence for Chemical Inference. II.[1] Interpretation of Low-Resolution Mass Spectra of Ketones[2]

A. M. Duffield, A. V. Robertson, Carl Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg

Contribution from the Departments of Chemistry, Computer Science, and Genetics, Stanford University, Stanford, California 94305. *Received November 18, 1968*

Abstract: A general approach to computer interpretation of the mass spectra of aliphatic ketones is described. Given the low-resolution mass spectrum plus the composition of the molecular ion, the computer program decides upon the most probable structure(s) consistent with the unknown's mass spectrum. The program makes extensive use of the DENDRAL algorithm.[1]

The application of digital computers to data reduction problems in both high-[3-6] and low-resolution[7] mass spectrometry has automated the accumulation of experimental information, especially in those instances where a gas chromatograph is directly interfaced with a mass spectrometer.[7] Interpreting this vast amount of data is a formidable problem for the research worker. Its solution may ultimately reside in the ability of suitably programmed computers to present a detailed interpretation of any mass spectrum. Preliminary results from other laboratories[8-11] have described methods whereby a computer participates in the interpretation of high-[8-10] and low-resolution[11] mass spectra. In addition, specific programs have been written for the determination of the amino acid sequence in small peptides from a computer interpretation of their high-resolution mass spectra.[12-14]

Another important contribution of computers in the identification of unknown mass spectra is the development[15] of information retrieval systems where files of mass spectra of known compounds are available for immediate comparison with spectra generated in the research laboratory.

We wish to describe in the present communication a general approach directed toward the complete computer interpretation of low-resolution mass spectra.[16] Our approach is based on the capability of a computer program (DENDRAL)[17] to manipulate structural representations of organic molecules and their functional groups and to generate rigorously exhaustive and irredundant lists of structures including the candidates for a given problem.[18] As far as we know, this general capability has not been embodied in other work attacking this problem. Aliphatic ketones were chosen as a suitable class of compounds with which to begin a computer interpretation of mass spectra in view of the fundamental knowledge available on their mass spectrometric fragmentation modes.[19]

The basic approach we have used in the computer interpretation of mass spectra is diagramatically represented in Scheme I. The DENDRAL program can enumerate all the possible acyclic chemical structures[1,18] of a given empirical formula. The program is fed a low-resolution mass spectrum of an unknown compound (including any metastable transitions observed) and the empirical composition of the molecular ion.[20] It then employs a theory of mass spectral fragmentation processes,[19] stored in the PRELIMINARY INFERENCE MAKER,[21] to decide what functional groups are present within the molecular structure of the unknown. The program now gives the inferred

(1) Part I: J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, *J. Am. Chem. Soc.*, **91**, 2973 (1969).

(2) This research was financed by the Advanced Research Projects Agency of the Office of the Secretary of Defense (Grant SD-183), the National Aeronautics and Space Administration (NGR-05-020-004), and the National Institutes of Health (Grants GM-11309 and AM-04257). The award of a Fulbright Travel Grant (to A. V. R.) from the Australian-American Educational Foundation is gratefully acknowledged.

(3) K. Biemann, P. Bommer, and D. M. Desiderio, *Tetrahedron Letters*, 1725 (1964).

(4) C. Merritt, Jr., P. Issenberg, M. L. Bazinet, B. N. Green, T. O. Merron, and J. G. Murray, *Anal. Chem.*, **37**, 1037 (1965).

(5) R. Venkataraghavan, F. W. McLafferty, and J. W. Amy, *ibid.*, **39**, 178 (1967).

(6) A. L. Burlingame, D. H. Smith, and R. W. Olsen, *ibid.*, **40**, 13 (1968).

(7) R. A. Hites and K. Biemann, *ibid.*, **39**, 965 (1967); **40**, 1217 (1968).

(8) (a) K. Biemann and P. V. Fennessey, 14th Annual Conference on Mass Spectrometry, Dallas, Texas, May 1966, p 322; (b) A. Mandelbaum, P. V. Fennessey, and K. Biemann, 15th Annual Conference on Mass Spectrometry, Denver, Colo., May 1967, p 111.

(9) R. Venkataraghavan and F. W. McLafferty, ref 8b, p 98.

(10) R. Venkataraghavan, G. E. Van Lear, and F. W. McLafferty, 16th Annual Conference on Mass Spectrometry, Pittsburgh, Pa., May 1968, p 139.

(11) B. Pettersson and R. Ryhage, *Arkiv Kemi*, **26**, 293 (1967).

(12) M. Senn, R. Venkataraghavan, and F. W. McLafferty, *J. Am. Chem. Soc.*, **88**, 5593 (1966).

(13) K. Biemann, C. Cone, B. R. Webster, and G. P. Arsenault, *ibid.*, **88**, 5598 (1966).

(14) M. Barber, P. Powers, P. Wallington, and W. A. Wolstenholme, *Nature*, **212**, 784 (1966).

(15) L. R. Crawford and J. D. Morrison, *Anal. Chem.*, **40**, 1464, 1469 (1968); S. Abrahamsson, *Science Tools*, **14**, 29 (1967).

(16) Although the program is currently restricted to the utilization of low-resolution mass spectra, no basic problems are anticipated in the future use of their complete high-resolution counterparts.

(17) J. Lederberg, in preparation.

(18) The programming of cyclic structures within DENDRAL is partially complete.[17]

(19) See for instance H. Budzikiewicz, C. Djerassi, and D. H. Williams, "Mass Spectrometry of Organic Compounds," Holden-Day, Inc., San Francisco, Calif., 1967; F. W. McLafferty, "Interpretation of Mass Spectra," W. A. Benjamin, Inc., New York, N. Y., 1967; K. Biemann, "Mass Spectrometry," McGraw-Hill Book Co., Inc., New York, N. Y., 1962.

(20) The determination of molecular formulas is a distinct problem assumed by DENDRAL to have been solved by other analytical approaches. At worst, a subroutine (the "change-making" algorithm) is called to produce a list of hypothetical molecular formulas compatible with the mass number of the molecular ion (J. Lederberg, "Computation of Molecular Formulas for Mass Spectrometry," Holden-Day, Inc., San Francisco, Calif., 1964).

(21) Program MODULES are labeled in upper case.

[Reprinted from the Journal of the American Chemical Society, 91, 2977 (1969).]

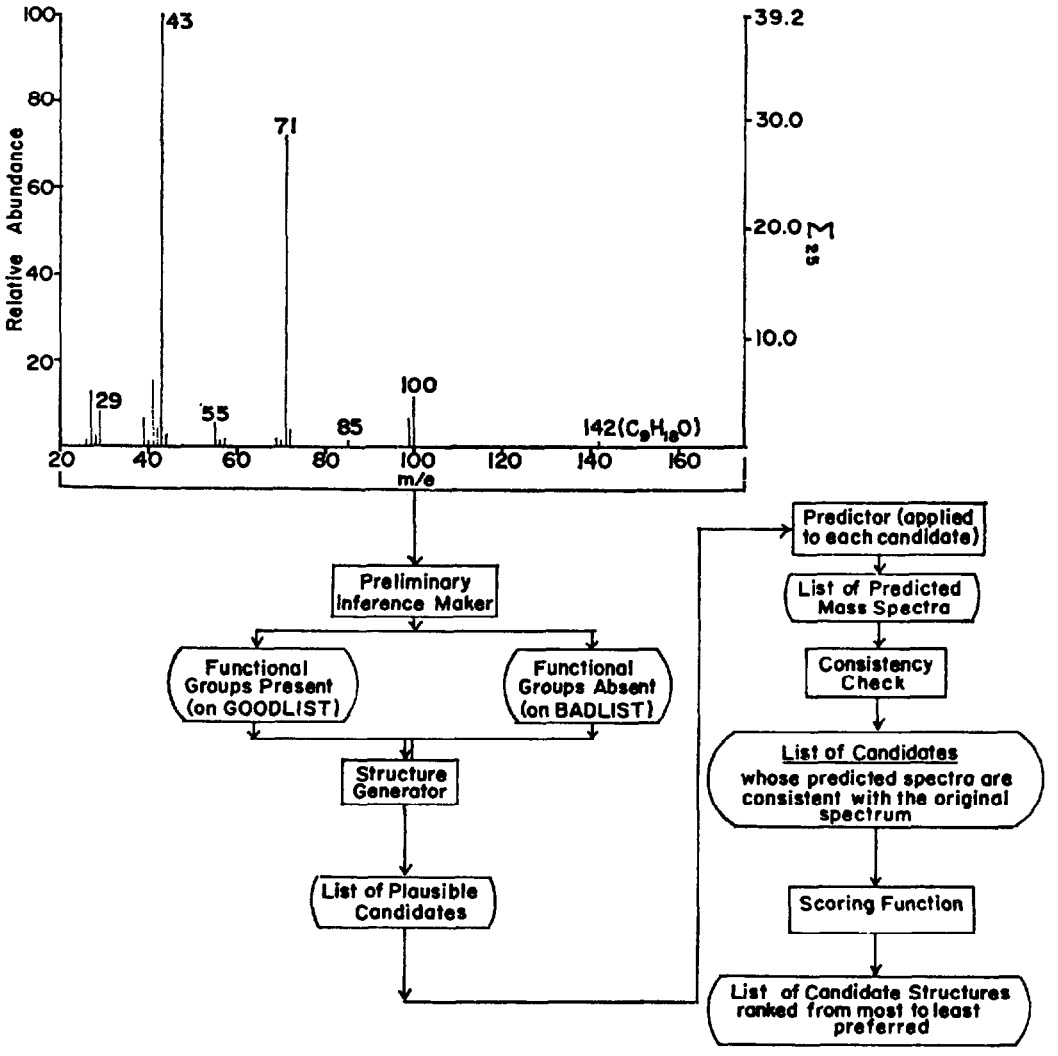


Figure 1. Mass spectrum of unknown aliphatic ketone and the conceptualization of Heuristic DENDRAL.

functional group highest priority, by placing it on GOODLIST. Other variants of the molecular configuration are placed on BADLIST and are not further considered. The STRUCTURE GENERATOR, using the DENDRAL algorithm,[17] then builds all possible candidate molecules within the constraints inferred from the spectrum. Each of these structures is successively scrutinized by the PREDICTOR which deduces the significant peaks of a hypothetical mass spectrum for each candidate structure. The program either rejects or accepts any structure on the basis of a comparison of the known and predicted mass spectra. The admissable molecules are then arranged in an ordered list by the SCORING FUNCTION.

This approach also summarized in Figure 1, is best understood by describing each major step of the program for a typical example using the low-resolution mass spectrum of an aliphatic ketone of composition $C_9H_{18}O$.

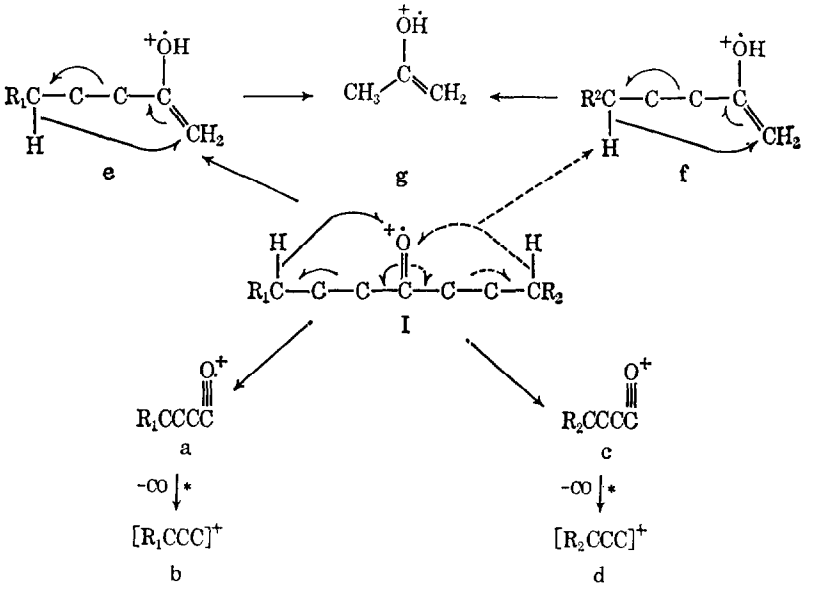
The first step is to identify the type of molecule represented by the unknown mass spectrum (Figure 1). The empirical composition $C_9H_{18}O$ could represent the molecular ion of either an aldehyde, ketone, or unsaturated aliphatic alcohol or ether (cyclic structures are deferred for the present).[18] With no constraints except a theory of instability[1] the STRUCTURE GENERATOR computed 1936 possible acyclic structures[1] corresponding to this composition. The first task of the program is to trim this number to a managable level.

This is accomplished within the PRELIMINARY INFERENCE MAKER by drawing on the accumulated practical experience of organic mass spectroscopists.[19] For instance, ketone mass spectra can be recognized because of the following simple fragmentation modes (Scheme I). Cleavage adjacent to oxygen, followed by decarbonylation of the product, yields two ion fragmentation pathways: I → a → b and I → c → d depicted in Scheme I. The McLafferty rearrangement ions e and f can yield the second rearrangement ion g provided the appropriate γ-hydrogen is present. A simple mathematical relationship then exists between the masses of the ions a and c and the molecular ion M, *viz.*

$$a + c = M + 28.$$

To identify ketones the PRELIMINARY INFERENCE MAKER searches for two significant peaks in the spectrum which satisfy this relationship. If more than one such pair exists, the program considers that pair of peaks which are present in the highest (combined) abundance. Next the program must recognize the decarbonylation products (whose abundance must be in excess of 10%

Scheme I



relative abundance) b and d and if metastable information is included each of these ions must be formed in association with a metastable ion.[22]

On completing an examination of an unknown spectrum with reference to these ketone rules, even if a ketone is recognized to be present, the program then searches for the possible presence of aldehydes, ethers, or alcohols according to a preselected list of conditions. For the sake of brevity the presentation of these constraints will be left to ensuing publications.

Since 82 of the 1936 possible isomers of $C_9H_{18}O$ are ketones, additional truncation is now desirable. The program searches for additional subgraphs after it has inferred the ketone group. The following four subgraphs are easily identified by the recognition of ions derived from the McLafferty rearrangement.[23] These have masses 58, 72, 86, 86, and 58 corresponding to the ions i, j, k, m, and g, respectively. Any substructure thus found is placed on GOODLIST and only those chemical graphs which contain this unit are constructed by the STRUCTURE GENERATOR. Should the program fail to recognize the mass number corresponding to the units represented by II–V (Scheme II), then these units are placed on BADLIST and all ketones *lacking* these subgraphs are generated.

Upon typing the sentence[24]

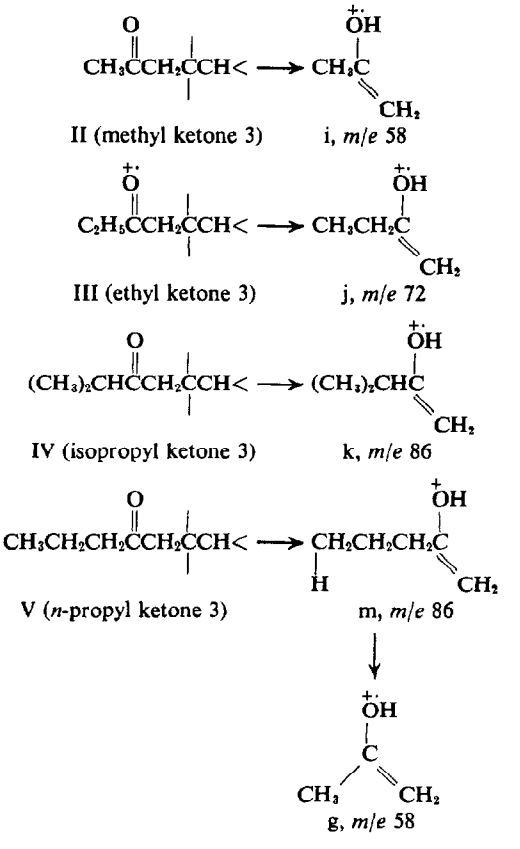
```
(EXPLAIN (QUOTE C9H18O) S:09490 (QUOTE TEST1) (QUOTE JUL-30-68))
```

the program within the PRELIMINARY INFERENCE MAKER responds with the following output.[25]

```
*GOODLIST = (*KETONE*)
*BADLIST = (*C-2-ALCOHOL* *PRIMARY-ALCOHOL* *ETHYL-ETHER2* *METHYL-ETHER2* *ETHER2* *ALDEHYDE* *ALCOHOL* *ISO-PROPYL-KETONE3* *N-PROPYL-KETONE3* *ETHYL-KETONE3* *METHYL-KETONE3*)
*PARTITIONS = (71. 43.)
```

Thus all possible acyclic ketones of composition $C_9H_{18}O$

Scheme II



will be constructed by the STRUCTURE GENERATOR excluding those which contain the subgraphs represented by II–V. In addition the program has been instructed to recognize the ions b and d and to construct only those ketones containing these groups of atoms on either side of the carbonyl group. In the present example this means those compounds containing C_3H_7 and C_5H_{11} flanking the keto group. Of the 82 possible ketones of composition $C_9H_{18}O$, 19 are propylpentyl ketones and the following eight candidate structures[26] are produced by

(22) It is our experience that the processes represented by a → b and c → d in aliphatic ketones are always accompanied by the presence of a metastable ion: see W. Carpenter, A. M. Duffield, and C. Djerassi, *J. Am. Chem. Soc.*, 89, 6167 (1967).

(23) F. W. McLafferty, *Anal. Chem.*, 31, 82 (1959).

(24) S:09490 is the code number under which Figure 1 is listed.

(25) The GOODLIST and BADLIST terminology refers to the following subgraphs: Ketone (–CO–), C-2 alcohol (>C(OH)CH_3), primary alcohol (–CH_2OH), isopropyl ketone 3 (*i*-PrCOCH_2CH<), *n*-propyl ketone 3 (*n*-PrCOCH_2CH<), ethyl ketone 3 (C_2H_5COCH_2CH<), methyl ketone 3 (CH_3COCH_2CH<), aldehyde (–OCH), alcohol (>COH), ethyl ether-2 ($C_2H_5OCH_2$–), methyl ether-2 (CH_3OCH_2–), ether-2 (–CH_2OCH_2–).

(26) It may help the reader to interpret DENDRAL dot notation if we name at least the first two entries. These are 5-methyloctan-4-one and 2,3-dimethylheptan-4-one, respectively.

the STRUCTURE GENERATOR in DENDRAL dot notation (in noncanonical order). Each of these remaining eight

```
C8*KETONE*H18
MOLECULES        NO DOUBLE BOND EQUIVS
  1.    <  CH..CH3 C3H7    C=.O C3H7 .
  2.    <  CH..CH3 CH..CH3 CH3    C=.O C3H7 .
  3.    <  CH..C2H5 C2H5    C=.O C3H7 .
  4.    <  C...CH3 CH3 C2H5    C=.O C3H7 .
  5.    <  CH..CH3 C3H7    C=.O CH..CH3 CH3 .
  6.    <  CH..CH3 CH..CH3 CH3    C=.O CH..CH3 CH3 .
  7.    <  CH..C2H5 C2H5    C=.O CH..CH3 CH3 .
  8.    <  C...CH3 CH3 C2H5    C=.O CH..CH3 CH3 .

DONE
```

ketones is now examined in detail to eliminate some and rank the rest. The PREDICTOR subroutine predicts an abbreviated low-resolution mass spectrum for each candidate structure which is then checked for possible inconsistencies with the original spectrum. Structures which show such inconsistencies are eliminated; finally, the SCORING FUNCTION ranks the remaining candidates. For instance the first candidate, 5-methyloctan-4-one, is eliminated since one can expect to observe a double McLafferty rearrangement ion[27] of mass 72, whereas the program recognizes that the original spectrum (Figure 1) contains only the ^{13}C isotope peak from *m/e* 71 at this mass value. This incompatibility between the predicted mass spectrum and Figure 1 prompts the following response by the computer.[28]

```
(SCORE (QUOTE TEST1) S:89498 )
JUL-30-68
1.)  *1C11CC1C1CC210C1C1CS

((43. . 44.) (71. . 100.) (72. . 13.) (99. . 55.) (100. . 13.) (142.
. 9.))
*THIS CANDIDATE IS REJECTED BECAUSE OF (IS072) .
```

Candidate 2, 2,3-dimethylheptan-4-one, from the output of the STRUCTURE GENERATOR is summarily eliminated by the program because of the disagreement between its predicted mass spectrum and Figure 1 while entries 3 and 4 can also be excluded from further consideration.

```
2.)  *1C11CC11CCC210C1C1CS

((43. . 44.) (71. . 100.) (72. . 4.) (99. . 55.) (100. . 2.) (114. .
2.) (142. . 9.))
*THIS CANDIDATE IS REJECTED BECAUSE OF (IS072 IS072 114.) .

3.)  *1C11C1CC1CC210C1C1CS

((43. . 44.) (71. . 100.) (86. . 4.) (99. . 55.) (114. . 4.) (142. .
9.))
*THIS CANDIDATE IS REJECTED BECAUSE OF (86. 114. 86. 114.) .

4.)  *1C111CCC1CC210C1C1CS

((43. . 42.) (71. . 100.) (86. . 4.) (99. . 57.) (114. . 4.) (142. .
9.))
*THIS CANDIDATE IS REJECTED BECAUSE OF (86. 114. 86. 114.) .
```

Candidates 5 and 6 showed no anomalous predicted peaks when compared to Figure 1, and thus remained viable candidates.

```
5.)  *1C11CC1C1CC210C11CCS

((43. . 58.) (71. . 100.) (99. . 58.) (100. . 15.) (142. . 18.))

6.)  *1C11CC11CCC210C11CCS

((43. . 58.) (71. . 100.) (99. . 58.) (100. . 2.) (142. . 18.))
```

The consistency check eliminates the final two compounds suggested by the STRUCTURE GENERATOR since both would be expected to expel 28 mass units (ethylene) by the operation of the McLafferty rearrangement process[23] and this fragmentation is absent (no peak at *m/e* 114) from Figure 1.

```
7.)  *1C11C1CC1CC210C11CCS

((43. . 58.) (71. . 100.) (99. . 58.) (114. . 2.) (142. . 18.))
*THIS CANDIDATE IS REJECTED BECAUSE OF (114.) .

8.)  *1C111CCC1CC210C11CCS

((43. . 48.) (71. . 100.) (99. . 51.) (114. . 2.) (142. . 18.))
*THIS CANDIDATE IS REJECTED BECAUSE OF (114.) .
```

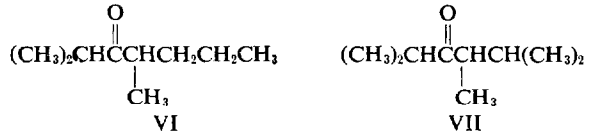
At this stage the program has been able to truncate the list of eight candidates (out of 1936 possible acyclic isomers of composition $C_9H_{18}O$ containing 82 ketones) to two candidates: 2,4-dimethylheptan-3-one (VI, entry 5) and 2,4,5-trimethylhexan-3-one (VII, entry 6). The SCORING FUNCTION was unable to distinguish between these two structures and ranked both as equally plausible. In point of fact the low-resolution mass spectrum, depicted as Figure 1, corresponds to VI.

```
*LIST OF RANKED MOLECULES:

1.  #5.
 S = 5.
  P = (43. 99. 71. 71. 100.)
  U = NIL

2.  #6.
 S = 5.
  P = (43. 99. 71. 71. 100.)
  U = NIL

*   1.  #N MEANS THE FIRST RANKED MOLECULE IS THE NTH IN THE
ORIGINAL NUMBERED LIST ABOVE. S = THE SCORE (HIGHEST = BEST) BASED
ON THE NUMBER OF SIGNIFICANT PREDICTED PEAKS IN THE ORIGINAL GRAPH.
P = THE LIST OF SIGNIFICANT PREDICTED PEAKS.  U = THE POSSIBLY
SIGNIFICANT UNRECORDED PEAKS USED TO RESOLVE SCORING TIES (THE FEWER
  IN DOUBT THE BETTER).
DONE
```



VI VII

Table I summarizes the program's interpretations of mass spectra for several aliphatic ketones, including the example just discussed in detail. If the SCORING FUNCTION ranks two or more candidates equally, as in this example, all are shown.

Heuristic DENDRAL was able to interpret the low-resolution mass spectra of the ketones listed in Table I by applying the rules enunciated earlier in this paper, *viz.*, identification of the masses corresponding to the ions a–d and any McLafferty rearrangement ions present. In addition the program was instructed to search for the presence of the "McLafferty plus one" rearrangement ion.[29] To do this it was necessary for Heuristic DENDRAL to calculate the ^{13}C-isotope contribution of the McLafferty rearrangement species and to subtract this value from the abundance of the ion one mass unit higher. Furthermore the presence of the "McLafferty plus one" rearrangement ion was taken as evidence in favor of a linear chain of at least five carbon atoms in the unknown ketone. This rule was responsible for the correct assignment of structure to the mass spectra of 3-octanone, 4-octanone, and 3-nonanone. However, it must be admitted that this rule failed in the case of 2-methyloctan-3-one as this compound lacked a peak in its mass spectrum attributable to the "McLafferty plus one" rearrangement.

The elimination of all four candidate structures proposed by the STRUCTURE GENERATOR in the example of 4-nonanone merits a brief statement. As noted in

(27) H. Budzikiewicz, C. Djerassi, and D. H. Williams, "Interpretation of Mass Spectra of Organic Compounds," Holden-Day, Inc., San Francisco, Calif., 1964, p 7.

(28) The first line of this response is a transliteration of line number 1 of the STRUCTURE GENERATOR'S output shown above. Next is the predicted list of mass-intensity pairs for this structure. The reasons for eliminating the candidate appear last.

(29) W. Carpenter, A. M. Duffield, and C. Djerassi, *J. Am. Chem. Soc.*, **90**, 160 (1968).

Table I. Heuristic DENDRAL's Interpretation of the Mass Spectra of Some Aliphatic Ketones

Compound	No. of aliphatic Isomers[a]	Ketones	No. of candidates from Structure generator	Consistency check	Ranking of candidates
2-Butanone[b]	11	1	1	1	1st, 2-butanone
3-Pentanone[b]	14	3	1	1	1st, 3-pentanone
3-Hexanone[c]	91	6	1	1	1st, 3-hexanone
2-Methylhexan-3-one[c]	254	15	1	1	1st, 2-methylhexan-3-one
3-Heptanone[b]	254	15	2	2	Tie for 1st, 3-heptanone and 5-methylhexan-3-one
3-Octanone[b,c]	698	33	4	4	1st, 3-octanone
4-Octanone[c]	698	33	2	1	1st, 4-octanone
2,4-Dimethylhexan-3-one[c]	698	33	4	3	Tie for 1st, 2,4-dimethylflexan-3-one and 2,2-dimethylhexan-3-one
6-Methylheptan-3-one[b]	698	33	4	4	1st, 3-octanone; tied for 2nd, 6-methylheptan-3-one, 5-methylheptan-3-one, and 5,5-dimethylhexan-3-one
3-Nonanone[c]	1936	82	7	7	1st, 3-nonanone
2-Methyloctan-3-one[c]	1936	82	4	3	Consistency check eliminated correct structure because no McLafferty + 1 peak was present in original mass spectrum
4-Nonanone[c]	1936	82	4	0	Consistency check eliminated all candidates since no peak was present at *m/e* 114 (McLafferty rearrangement) in original mass spectrum

[a] These numbers were computed by Heuristic DENDRAL operating within the constraints of chemical stability (BADLIST) described in ref 1, or were extrapolated by hand from semilog plots of the tables generated by the program.[1] [b] Literature mass spectrum: A. G. Sharkey, J. L. Shultz, and R. A. Friedel, *Anal. Chem.*, **28**, 934 (1956). [c] Mass spectrum determined at Stanford University, Chemistry Department.

Table I all four structures were rejected because of the absence of a peak at *m/e* 114 (M − 28) due to a McLafferty rearrangement from the *n*-propyl side chain in the mass spectrum of 4-nonanone. To overcome this failure Heuristic DENDRAL has been instructed that when two McLafferty rearrangements are possible in an aliphatic ketone with one from a linear chain of more than five carbon atoms and the other involving transfer of a primary hydrogen atom, then the latter rearrangement may be absent.

In conclusion we believe that although the present capability of Heuristic DENDRAL for the interpretation of unknown mass spectra is limited, this approach definitely merits further attention, especially when additional selected spectroscopic data (infrared, ultraviolet, and nuclear magnetic resonance spectra) are included.

Experimental Section

The computer program described here, named Heuristic DENDRAL, runs on the PDP-6 time-sharing computer at the Stanford University Artificial Intelligence Laboratory. It is written in the LISP programming language in two large parts, shown as the left- and right-hand sides of Figure 1. Each part requires approximately 40 K of core memory, although there is an estimated 15–20 K of overlap between the two parts. Although many factors influence the length of time the program takes from the time it receives the initial spectrum and molecular ion composition to the time it outputs its ordered list of explanatory structures, 4 or 5 min at the console will usually suffice for examples of the size described here. The greatest amount of time is used by the PREDICTOR; thus our current efforts are concentrated on reducing the list of plausible candidates by giving more theoretical information to the PRELIMINARY INFERENCE MAKER.

The program is now confined to monofunctional acyclic structures. However, we are currently working on removing these fundamental limitations as well as on adding more mass spectrometry theory to the PREDICTOR and the PRELIMINARY INFERENCE MAKER.